SynPharm: a Guide to PHARMACOLOGY database tool for designing drug control into engineered proteins

ACS OMEGA

Article

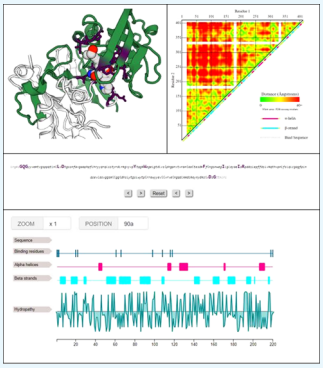
Cite This: ACS Omega 2018, 3, 7993−8002

SynPharm: A Guide to PHARMACOLOGY Database Tool for Designing Drug Control into Engineered Proteins

Sam M. Ireland,[†] Christopher Southan,* Alazne Dominguez-Monedero, Simon D. Harding, Joanna L. Sharman, and Jamie A. Davies

Centre for Discovery Brain Sciences, Deanery for Biomedical Sciences, University of Edinburgh, Edinburgh EH8 9XD, U.K.

ABSTRACT: A major challenge in synthetic biology, particularly for mammalian systems, is the inclusion of adequate external control for the synthetic system activities. Control at the transcriptional level can be achieved by adaptation of bacterial repressor−operator systems (e.g., TetR), but altering the activity of a protein by controlling transcription is indirect and for longer half-life mRNAs, decreasing activity this way can be inconveniently slow. Where possible, direct modulation of protein activity by soluble ligands has many advantages, including rapid action. Decades of drug discovery and pharmacological research have uncovered detailed information on the interactions between large numbers of small molecules and their primary protein targets (as well as off-target secondary interactions), many of which have been well studied in mammals, including humans. In principle, this accumulated knowledge would be a powerful resource for synthetic biology. Here, we present SynPharm, a tool that draws together information from the pharmacological database GtoPdb and the structural database, PDB, to help synthetic biologists identify ligand-binding domains of natural proteins. Consequently, as sequence cassettes, these may be suitable for building into engineered proteins to confer small-molecule modulation on them. The tool has ancillary utilities which include assessing contact changes among different ligands in the same protein, predicting possible effects of genetic variants on binding residues, and insights into ligand cross-reactivity among species.



INTRODUCTION

Synthetic biology is a technology for engineering of new biological functions through the construction of novel genetic networks to realize novel metabolic, signaling, and developmental pathways.[1−5] Some synthetic biological systems use only natural proteins (i.e., as represented by the Swiss-Prot canonical sequence for that species) or achieve novel functions by combining proteins not normally found in the same cell or even the same species. Other systems involve the use of novel proteins, themselves typically including domains chosen from various natural proteins and coded into an engineered gene: an example is the SynNotch synthetic cell−cell signaling system.[6] In many applications, there is a clear need for synthetic biological devices to be subject to external controls, for example, to create adequate safeguards and to exert temporal and/or spatial control on a particular system. This need is particularly acute when the device is intended to be used in the general environment or in a medical implant. At the very least, there needs to be a reliable means to shut the system down quickly, and much thought is being given to this problem.[7,8]

Most control systems used to date have operated by using small molecules to control gene transcription. Typically, they use antibiotic-sensitive transcriptional repressor proteins from bacterial systems, the operator sites of which are fused to the promoter of the synthetic gene: the well-known tetR system is a much-used example.[9] These systems work well but their effect on protein activity is very indirect, blocking transcription of further mRNA for a protein but not affecting existing molecules of the protein itself nor of the mRNA from which new protein molecules will be translated. Constitutive differences in mRNA half-lives can, however, limit this approach for particular proteins.[10] Direct control of protein activity would be faster, which is why this dominates natural inter- and intracellular signaling. For synthetic circuits, control by rapidly diffusing small molecules would be particularly useful and several novel controls of this type have been constructed, generally by a laborious process of selection from large libraries of protein variants.[11,12]

As modulators of activity, small molecules have many advantages over alternative forms of experimental functional modulation, such as CRISPR, RNAi, and antibody blocking. Principal advantages of these are as follows: (a) rapid action; (b) dose response can be used to vary the effects quantitatively; (c) reversal by wash-out; (d) use of equal and low-potency analogues with different chemotypes as specificity and reproducibility controls; (e) although less common, activators or agonists may be suitable for positive modulation (i.e., gain-of-function interventions); (f) allosteric modulators offer a different type of kinetic control; and (g) small molecules can be accurately measured both pre and post experiment (e.g., to monitor input dosing and metabolic degradation).

Received: April 5, 2018
Accepted: June 26, 2018
Published: July 18, 2018

The need for pharmacological researchers to access data for the interactions between druglike molecules and their protein targets has resulted in the production of a range of databases aligned to this general task, starting with BindingDB in 2001.[13] Updates on these resources have recently been reviewed.[14] These databases present valuable sources of information that might help synthetic biologists identify drug−protein pairs in which the drug-binding site of the protein is small and self-contained enough to be used as a "module" that will confer drug control on engineered proteins. This would allow rapid and direct modulation of the activity of the protein without the lag times involved in transcription, translation, and degradation. The use of an approved drug as the controlling ligand would bring the additional advantage in that safety aspects of clinical drugs, and their possible off-target side effects, are generally well established. This makes the approach particularly valuable if the synthetic biological constructs are eventually to be translated into in vivo, clinical, or animal-agricultural contexts. However, attempting such module selection from large-scale chemogenomic databases such as BindingDB,[15] ChEMBL,[16] and even directly from PDB[17] would be challenging. Various types of PDB abstractions such as the sc-PDB ligand-binding database[18] and PDBbind[19] are also useful resources but have long update cycles.

To make navigating these complex datasets easier, we have created a web-based tool that integrates pharmacological and protein-binding information as a first-stop entry point for the drug-binding domains of selected proteins in a manner useful to synthetic biologists. The interface we designed supports a variety of searching and browsing strategies and facilitates the choosing of the most appropriate protein domain to be used as a controllable module for a particular purpose. This functionality, that we have named SynPharm, has been integrated as a tool within the IUPHAR/BPS Guide to PHARMACOLOGY database (GtoPdb), an expert-curated, open-access database by the International Union of Basic and Clinical Pharmacology.[20] This was chosen for the following reasons:

1. It is embedded in an environment with an active experimental synthetic biology team. This means that the initial bioinformatics in vitro testing cycles are already in progress (and the latter will feed back to enhancements of the former).
2. GtoPdb has a relatively rapid release cycle of approximately 2 months, and it is intended to synchronize SynPharm updates.
3. Relative to the larger resources our less broad-ranging but pharmacologically selective PDB mappings present much smaller sets for users to easily navigate but still capture approved drugs and clinical candidates.
4. Every ligand in SynPharm is expert-curated and activity-mapped even though this activity is not always explicitly referenced in the publication associated with the PDB entry.
5. This means that our selected ligands are also manually identified as authentically binding to specific protein pockets rather than inorganic ions and/or heteroatoms from crystallization reagents.
6. Partially due to SynPharm but also because of the increasing interest in new receptor and enzyme ligand structures in general, we have been recently enhancing our capture by triaging all new human PDB depositions.
7. Beyond direct application to synthetic biology per se, SynPharm has ancillary utility for GtoPdb users to explore ligand structures.

The Results section below presents the web pages that we have instantiated for SynPharm, the technical construction of which is described in the "Methods" section.

RESULTS

Our ligand-identification process identified 804 ligand−target interactions that were associated with at least 1 PDB code. Manually checking these interaction—PDB maps and rejecting duplicates gave a preliminary list of 768 interactions with associated PDB files. Among the interactions with structural data, 744 of the 768 (97%) interactions concern human data, with 15 (2%) rat, 8 (1%) mouse targets, and 1 *Plasmodium falciparum* target. The statistics reported, including for the web page captured in figures, were distilled from GtoPdb release 2018.1. They will thus change with subsequent releases, mainly from the curation of new PDB ligands but also some cases where PDB structures with activity data against new targets are reported.

Our results established (not unexpectedly) that the distribution of interactions for which there is identifiable structural data is unevenly distributed among target families, as shown below in Table 1.

Table 1. Representations of Different Classes of Targets in GtoPdb That Have Any Interaction Data and That Have Useful Structural Data[a]

GtoPdb target type	targets with interactions (with or without structures)	number of interactions (with or without structures)	targets with structural data	interactions with structures
GPCR	277 (16%)	9078 (52%)	29 (12%)	67 (11%)
enzyme	755 (44%)	3518 (20%)	157 (64%)	365 (60%)
VGIC	127 (7.5%)	1408 (8%)	2 (0.8%)	3 (0.5%)
LGIC	66 (3.9%)	1027 (5.9%)	4 (1.6%)	4 (0.660%)
other ion channel	47 (2.8%)	201 (1.2%)	0	0
catalytic receptor	178 (10%)	992 (5.7%)	13 (5.3%)	40 (6.6%)
NHR	35 (2.1%)	523 (3.0%)	25 (10%)	104 (17%)
transporter	120 (7%)	433 (2.5%)	1 (0.4%)	4 (0.66%)
other protein	99 (5.8%)	231 (1.3%)	16 (6.5%)	23 (3.8%)

[a]GPCR = G protein-coupled receptors; VGIC = voltage-gated ion channels; LGIC = ligand-gated ion channels; NHR = nuclear hormone receptors.

As is well known, some target classes are more tractable to X-ray determination and consequently proportionally more highly represented with structural data. For example, nuclear hormone receptor (NHR) interactions are particularly structure-dense in comprising 17% of the annotated sequences but just 3% of GtoPdb interactions overall. Enzymes are also over-represented in that just 20% of GtoPdb interactions involve enzymes, but 60% of those proteins with structural data. By contrast, voltage-gated ion channels (VGICs) have just 3 annotated sequences (0.5%), compared with 1408 (8%) total interactions. This bias reflects the inherent difficulties with structural studies of membrane proteins, although recent advances have led to an increase in the number of GPCR

DOI: 10.1021/acsomega.8b00659
ACS Omega 2018, 3, 7993−8002

structures in the last few years, many of which include ligands.[21]

Our process of compiling the SynPharm resource, detailed in the Methods section, is outlined in Figure 1. The output has been used to populate the home page designed to allow users to search the dataset by ligand or target protein (Figure 2).

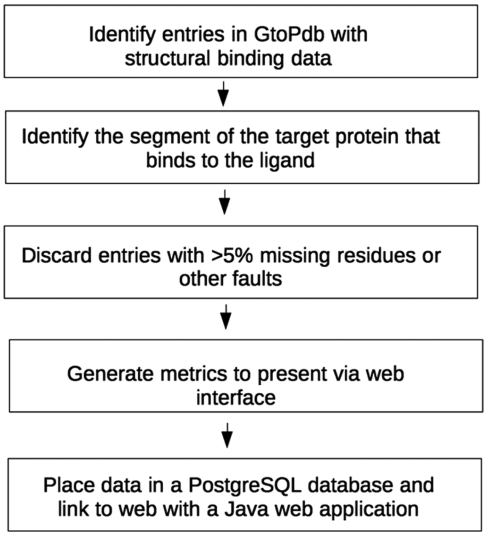


Figure 1. Strategy used to produce a database of potentially useful interactions from known binding data.

Users may browse the site without a specific molecule in mind or alternatively take as their starting point identifying any ligand-binding segment of a protein that might be transferrable to an engineered protein in their project. In this case, clicking on the "Sequences" link without entering a search term lists all target proteins in the list of potentially useful pairs described above. This list can be browsed as shown in Figure 3.

In Figure 3, targets have been ordered by the length of ligand-binding segment but they can be ordered by any of the columns by clicking on the table headers. These metrics can provide useful first-pass information to prioritize more detailed analyses. Selecting any target brings the user to its sequence page. At the head of this page is a three-dimensional visualization of the target chain bound in complex with the ligand, with the binding segment itself highlighted in green to show its context within the original chain (Figure 4A).

The views in Figure 4 provide a rapid visual indication of how independent of the other features of the protein the binding segment's structure is likely to be and thus more transferable to other proteins. Visualization of the structure uses the JavaScript PV protein viewer.[22] In addition to showing the structures, the sequence pages present metrics such as proportional chain length and contact ratio (used as a rough measure of likelihood that the sequence will fold correctly by itself, as it is a measure of "domain-likeness": higher is more promising). The GtoPdb affinity data for the specific ligand–target interaction are also provided. Each sequence also has a residue distance matrix (Figure 4B), which depicts the

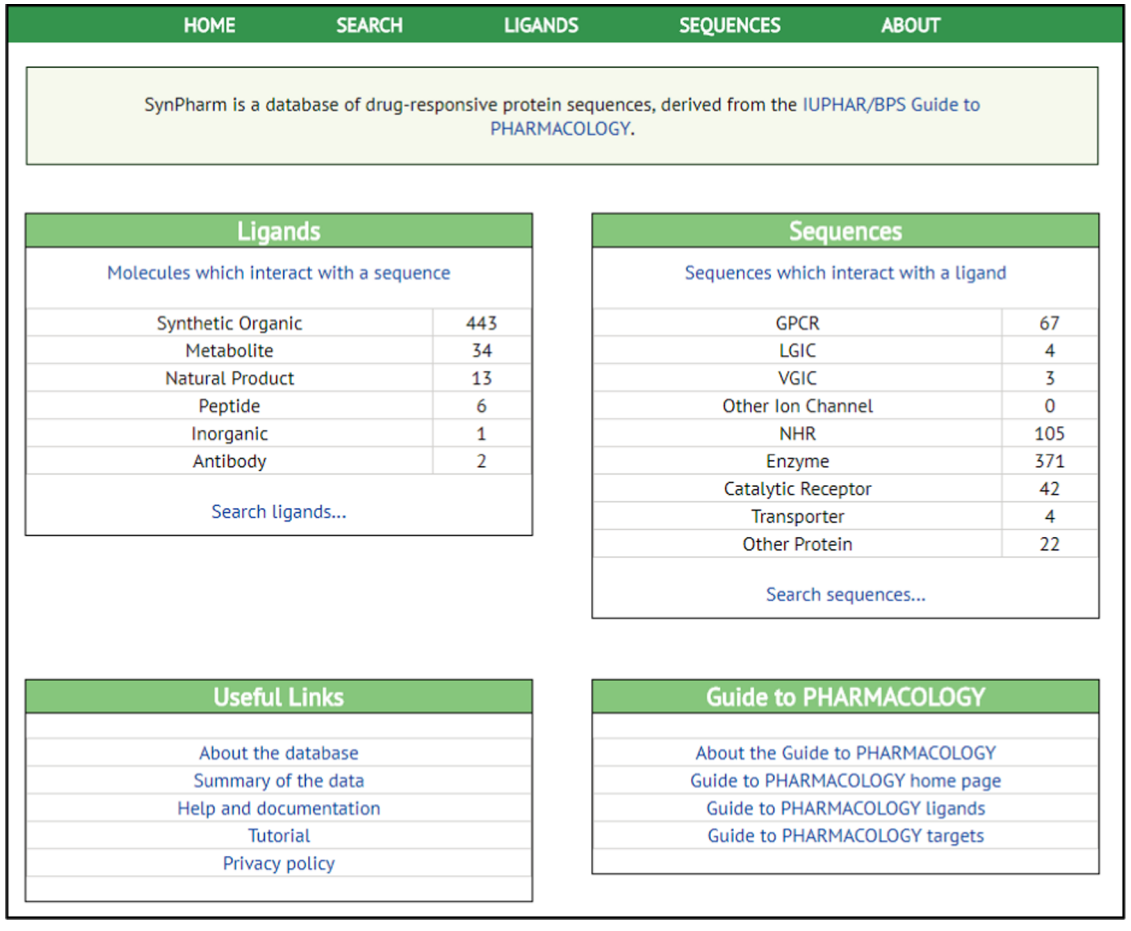


Figure 2. SynPharm home page with summary statistics at http://synpharm.guidetopharmacology.org/.

DOI: 10.1021/acsomega.8b00659
ACS Omega 2018, 3, 7993−8002

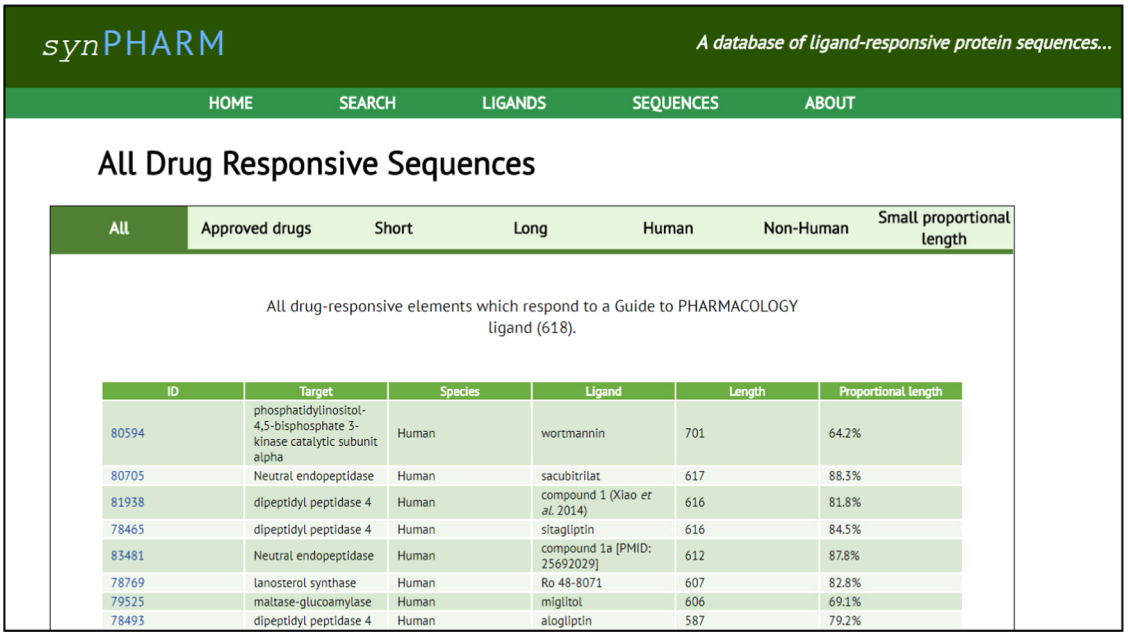


Figure 3. Top section of the list served to a user entering the target sequence part of the database. http://synpharm.guidetopharmacology.org/sequences/.

distances between any two given residues in the binding chain, with the bind sequence itself highlighted with a black dotted line. This is to give a sense of the globularity of the sequence within the chain, and how compact it is.

There is also a feature viewer (Figure 4D) for each sequence, which utilizes the biojs-vis-protein features viewer.[23] In addition to binding residues and secondary structure elements, the feature viewer also maps hydrophobicity along the bind sequence, using the Kyte−Doolittle measures of amino-acid hydrophobicity.[24] There are extensive search functions for identifying sequences or ligands by various metrics. All ligands have links back to GtoPdb, and a subset of their data is available directly on the SynPharm page, particularly molecular data and clinical approval information. These were chosen because they may be relevant to a researcher when picking a molecular switch inducer, but the full range of pharmacological data is accessible via the link back to GtoPdb. This can be illustrated for BACE1, an aspartyl protease drug target for Alzheimer's disease.[25] In Figure 5, a section of the BACE1 entry is shown and Figure 6 shows a sequence alignment of the extracted ligand interaction sections.

The eight SynPharm entries are indicated in the upper panel. From the total ligand entries in the lower panel, affinity values are displayed for three of the SynPharm ligands, compound 16 [PMID: 23412139], AZ-4217, and AMG-8718. The PDB ligands crystallized in a target are indicate with the red circular logo intersected with a helix (note that verubecestat did not pass the SynPharm fault filters but the PDB entry can still be inspected).

A number of SynPharm advantages can be discerned from the BACE1 example, particularly because it is one of the most intensively perused drug targets (reflecting the massive unmet need for effective Alzheimer's treatments). Metrics in support of this are that no less than 364 BACE1 structures are in PDB (nearly all with ligands), 11 of which were deposited in 2017. The ChEMBL 23 human BACE1 entry is linked to 6846 structures with some level of activity mapping. The GtoPdb BACE1 entry maps 21 quantitative ligand interactions with a focus on clinical candidates and stringently selected research leads. Of these, 11 are in PDB (indicated with the orange circular logo) and 5 are not in ChEMBL. From the 11, the 8 indicated in Figures 4 and 5 have passed the SynParm triage (described in the Methods section) and, as multiple ligands for the same target, provide a useful calibration.

For example, the alignments shown in Figure 5 indicate explicit differences among ligand interaction residues for the set even though the alignment of the sequence sections indicates they are binding to the same pocket. We can note that all eight interact with Tyr 132 whereas only 82636 interacts with Glu 134 and Gly 135 and Tyr 137 only interacts with 78987 and 84541. These differences may be spatially minor (i.e., possibly only just outside the 5 Å limit used by SynPharm) but can nonetheless be useful. Even more useful to the synthetic biologist is to compare the overall length of the contiguous binding section for particular ligands. In Figure 5, we can see that five sequences extend out to Ala 396 as the ultimate interaction point. However, the results also indicate that only extending to DSGTT (or just past it in cassette terms) may be sufficient.

Although SynPharm would be sufficient as a stand-alone tool, there are external resources that complement it. The most obvious of these are the primary data sources of RCSB PDB and PDBe, both of which both of which have complementary features for visualizing ligand binding in a sequence context. We would also recommend PDBSum for other types of display.[26] These include advanced two-dimensional secondary structure diagrams, the LIGPLOT display of ligand binding, and indications of sequence conservation. In cases where there are many ligands co-crystalized in the same protein (e.g., for BACE1), the PocketOme encyclopedia of small-molecule binding sites will give a detailed breakdown of ligand sets.[27] For a deep exploration of both sequence- and structure-based homology, we suggest the Phyre2 web portal for protein modeling, prediction, and analysis.[28] For ligands per se (with or without PDB entries), we have made another important utility accessible from within GtoPdb in the form of ChEMBL

DOI: 10.1021/acsomega.8b00659
ACS Omega 2018, 3, 7993−8002

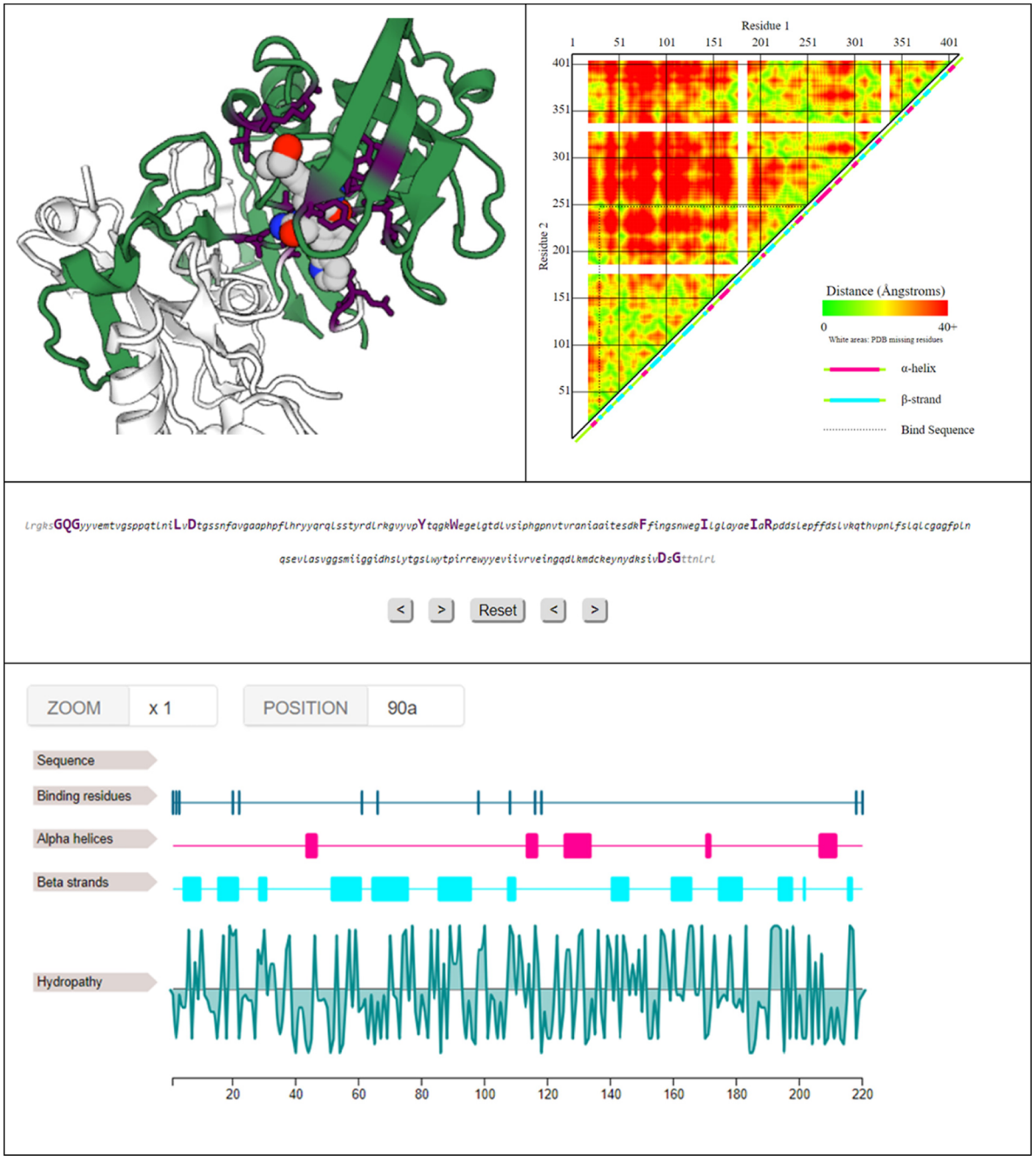


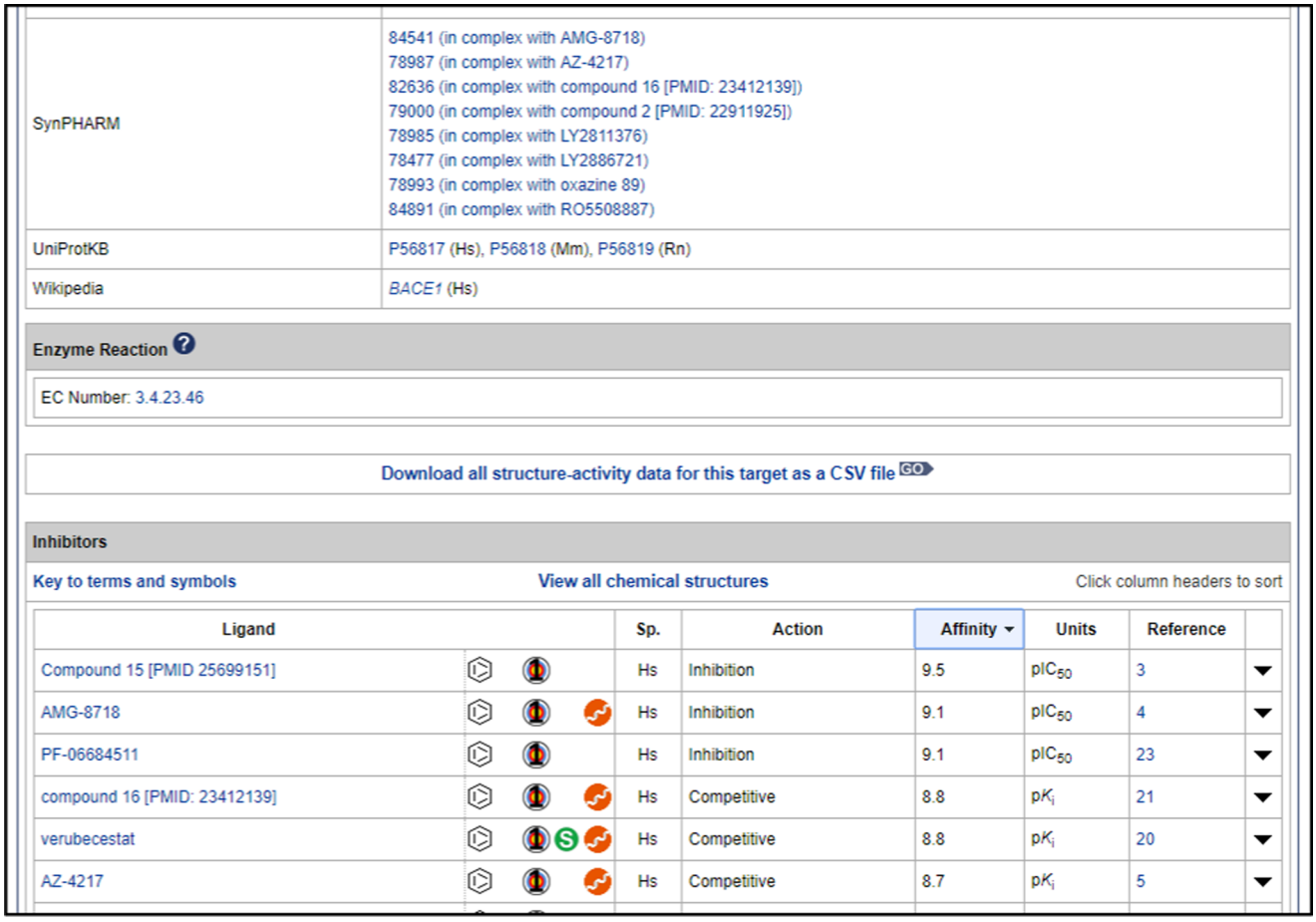
Figure 4. Examples of the types of structural display found on the sequence details page for human β-secretase 1 in complex with the ligand AMG-8718 (sequence ID 84541). The top-left panel shows the three-dimensional structure interactive viewer where the binding segment is highlighted in purple, the rest of the target protein in green, and the ligand is shown in stick view. The top-right panel shows the residue distance matrix. The distance between any two residues in the target chain is denoted by color, green to red, and, on desktop screens, hovering over any pixel will provide an exact numerical distance in angstroms of the relevant residues. White portions denote residues missing from the PDB file of origin. The dotted line indicates the binding sequence. The central panel shows the binding portion of the sequence. The arrows allow the sequence sections to be extended outward beyond the first and last interaction residues (five are shown on each end in this case) The lower panel shows a zoomed-in section of the feature viewer. Binding residues are shown in context with secondary structure elements (α-helices and β-strands) and hydrophobicity over the peptide sequence.

outlinks. For BACE1, this means that we were able to find one of the highest reported potencies for a lead compound with a 0.3 nM IC50 against the purified enzyme (ligand ID 9982, compound 15 [PMID 25699151]). This would be of interest to test against an engineered protein despite the absence of a PDB entry.

DISCUSSION

The work described in this paper has resulted in a new open web resource mainly designed to help synthetic biologists to engineer pharmacological regulation into their proteins. The idea of adding regulation into engineered proteins has already proved itself useful in a variety of contexts. A famous example is the addition of the tamoxifen-sensitive ERT2 domain into

DOI: 10.1021/acsomega.8b00659
ACS Omega 2018, 3, 7993−8002



SynPHARM	84541 (in complex with AMG-8718) 78987 (in complex with AZ-4217) 82636 (in complex with compound 16 [PMID: 23412139]) 79000 (in complex with compound 2 [PMID: 22911925]) 78985 (in complex with LY2811376) 78477 (in complex with LY2886721) 78993 (in complex with oxazine 89) 84891 (in complex with RO5508887)
UniProtKB	P56817 (Hs), P56818 (Mm), P56819 (Rn)
Wikipedia	*BACE1* (Hs)

Enzyme Reaction

EC Number: 3.4.23.46

Download all structure-activity data for this target as a CSV file GO

Inhibitors

Key to terms and symbols | View all chemical structures | Click column headers to sort

Ligand	Sp.	Action	Affinity	Units	Reference
Compound 15 [PMID 25699151]	Hs	Inhibition	9.5	pIC_{50}	3
AMG-8718	Hs	Inhibition	9.1	pIC_{50}	4
PF-06684511	Hs	Inhibition	9.1	pIC_{50}	23
compound 16 [PMID: 23412139]	Hs	Competitive	8.8	pK_i	21
verubecestat	Hs	Competitive	8.8	pK_i	20
AZ-4217	Hs	Competitive	8.7	pK_i	5

Figure 5. Snapshot from the GtoPdb BACE1 target entry, with ligands ranked by affinity values. http://www.guidetopharmacology.org/GRAC/ObjectDisplayForward?objectId=2330.

Cre recombinase to create a drug-inducible gene excision, allowing experimenters to remove gene function from an experimental animal at the time of their choosing.[29] This has been used in a variety of applications. Some have used the system to genetically mark cells for lineage tracing in development,[30,31] disease,[32] and regeneration.[3,33] Other applications have used ligand-inducible Cre to examine gene function by removing it from a cell only at a chosen stage of development.[34,35] Induced Cre-mediated recombination has also been used to create sarcomas in model animals for the purposes of studying tumor development.[36] The technique has even been used in anatomical studies, deliberately suboptimal doses of inducer being used to mark only sparse neuronal cells, allowing their detailed morphology to be studied in otherwise unlabeled tissue.[37] The use of photocaged estrogen adds an optogenetic dimension to a version of the system using Cre-ER instead of Cre-ERT2, allowing light to be used to activate Cre-ERT2 in specific cells.[38]

The addition of ligand control is not limited to Cre. A similar technique has been used to add the ER domain to Snail, to study the role of that mediator of epithelial−mesenchymal transitions in controlling fibrosis in adult kidney disease.[39] A recent example of a construct design success using SynParm is provided by our own work in placing the effectors of CRISPR, Cas9 and Cpf1, under the control of tamoxifen and mifepristone (Dominguez-Monedero et al., manuscript submitted). The impact of these examples establishes that engineering control into proteins can be useful. It is our hope that the tools described here will be useful in the construction of further examples, broadening the range of ligands that can be used for this type of work.

Several caveats should be taken into consideration with our approach. One of these is the necessary restriction to contiguous sections of sequence. However, it is well known that overall binding energies are likely to have at least some contribution from long-range secondary structure interactions within the entire protein structure. Thus, the binding sequence cassette not only needs to fold correctly within the engineered host sequence, but it may also have a lower binding constant and altered kinetics (e.g., K_{on} and/or K_{off}) compared with the full length native counterpart. This also means that the discontinuous binding sites characteristic of receptors, ion channels, and transporters are largely excluded from our data harvest (but the associated ligands are not necessary ruled out for synthetic applications). Another caveat is that GtoPdb literature selection focuses on clinical candidates where optimization often result in a lower potency than the initial lead compounds. This bias is thus not optimal for ligand-binding cassettes. Notwithstanding, for in vitro synthetic biology applications, complementary data can be explored, including searching for very potent inhibitors that are neither in GtoPdb nor in the PDB but have a high likelihood of binding the same sequence section (and this could be supported by structural superimposition and/or docking experiments). We note also the caveat that the nesting-in of active site sections, by definition, could endow the host protein

DOI: 10.1021/acsomega.8b00659
ACS Omega 2018, 3, 7993−8002

Figure 6. Differences in the contact residues extracted for the eight BACE1 ligands. The eight SynPharm sequences (in descending order) are 84891, 78993, 78985, 78477, 78900, 82636, 78987, and 84541. The latter (lowermost) is for AMG-8718 as shown in Figure 3. As for the SynPharm display, the uppercase letters indicate ligand contact residues.

with enzyme activity. In such cases, it should be possible to abolish such unwanted properties by mutating active site residues that are not major contributors to the binding energy. Alternatively, because GtoPdb has annotated a number of allosteric ligands, these noncatalytic binding modules could be exploited.

DOI: 10.1021/acsomega.8b00659
ACS Omega 2018, 3, 7993−8002

We can point out utilities of SynPharm that extend beyond practical applications to synthetic biology per se. First, the entries simply act as a convenient flag to users for the existence of relevant PDB structures, along with the orange logo. Second, there is increasing interest in the effects of protein sequence variants that affect protein function in pharmacologically significant ways, for example, patient drug responses if substitutions are found in the SynPharm sequences for individuals and population groups. Third, by adding rodent or other model organism sequences to the sequence alignments shown in Figure 5, insight can be gained into orthologous cross-reactivity of ligands that could be experimentally tested. An example for BACE1 is that the longest sequence section from Figure 5 has 82% identity with the Zebrafish orthologue (UniProt Q6NZT7). Although no structure of this protein is yet available, the SynPharm results indicate that there are differences in the vicinity of the binding residues. Notwithstanding, the similarity suggests that functional perturbations could be carried out (e.g., with compound 15 [PMID 25699151]) in this important model organism for human disease conditions. Matching ligand-binding sequences to distant homologs raises the possibility of predicting binding sites in proteins rather than relying on known ones. Although this goes beyond the functionality of SynPharm per se, this could be generally applicable in GtoPdb. Probable binding pockets of compounds with potent affinities may be predictable for human paralogues or species orthologues on the basis of homology modeling (e.g., using Phyre2[28]).

We would be pleased to hear from other teams who would like to use SynPharm, and we may be able to assist in cross-checking complementary sources to expedite their choices. In addition, we would be like to record future examples of success that we could reference.

METHODS

We used a sequential bioinformatic strategy for identifying ligand-binding sequence sections potentially useful to synthetic biology (Figure 1). Stage 1 was a screen for targets in GtoPdb for which any structural ligand-binding data were available in the form of RCSB PDB files. This screen was performed by using GtoPdb web services to request PDB codes for each ligand associated with a target in GtoPdb (2018.1 release). To obtain further structural data on this first set of potentially interesting interactions, the RCSB PDB web services were queried with information from GtoPdb. For each ligand−target interaction, PDB codes associated with ligands were obtained by searching on ligand code, name, SMILES, InChI, or peptide sequence. PDB codes associated with targets were obtained by searching the RCSB PDB web services using UniProtKB accessions.

Stage 2 was to identify amino-acid residues on the target that mediate each of the ligand−target binding interactions. The residues that mediate the ligand binding were identified by either using the information in REMARK 800 and SITE records of the relevant PDB file or, if no such records exist, by selecting all residues with atoms within 5 Å of a ligand atom (ignoring hydrogen atoms). The binding sequence was then defined as the segment of the protein chain that contained all the ligand-binding residues, for example, a segment between amino acids 30 and 45 of a protein chain. If binding residues were on more than one peptide chain of a multipeptide target protein complex, the interaction was rejected as not being useful for the purposes of protein engineering. Interactions were also rejected if more than 5% of the residues in the chain are "missing", that is, not observed in the PDB file (according to REMARK 465 records). This was the most frequent reason for discarding candidates. Stage 2 cut the list of potentially useful interactions down to 618 sequences. This is a relatively small proportion (3.5%) of the number of interactions in the 2018.2 release of GtoPdb, a reflection of the small number of PDB target−ligand interactions that pass our filtration rules for SynPharm.

Stage 3 associated certain metrics with each sequence. These were as follows: (i) its length as a proportion of the original chain, (ii) its "contact ratio"—defined as the ratio of internal contacts (all nonhydrogen atom pairs within the sequence within 5 Å of each other, excluding atoms within two covalent bonds of each other) to external contacts (all nonhydrogen atom pairs between the sequence and the rest of the chain, less than 5 Å). In cases where an interaction had multiple PDB maps and so multiple potential sequences to represent it, we selected those with the smallest length proportional to their original chain length as the most likely to be useful for engineering purposes. The system also allows manual selection of an interaction—PDB map if this is required.

The functions for accessing the GtoPdb web services have been bundled into a stand-alone Python library called pyGtoP,[40] and the code for parsing PDB files and identifying the various elements within them (used in sequence construction) has been bundled into a Python PDB parser called molecuPy. Both are open source projects viewable on GitHub. The scripts that used these new libraries to do the work described above, as well as the code for the database and web interface itself, are also open source and viewable on GitHub in the SynPharm repository.[41,42]

Construction of a Web Interface. Our aim was to make the data available in a useful format to synthetic biologists, in the form of an easy-to-use web page. We have therefore stored the data in a PostgreSQL[43] database, with a separate staging database to make future updates easier. This is connected to a web page using a Java (Oracle Corporation, Redwood City, CA) web application installed on an Apache Tomcat web server (The Apache Software Foundation), and the web page is open access at ref 44.

AUTHOR INFORMATION

Corresponding Author

*E-mail: cdsouthan@hotmail.com.

ORCID

Christopher Southan: 0000-0001-9580-0446

Alazne Dominguez-Monedero: 0000-0002-0146-0602

Present Address

†Biomolecular Structure & Modeling Unit, Institute of Structural and Molecular Biology, Division of Biosciences, University College London, London WC1E 6BT, U.K. (S.M.I.).

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

S.M.I. and A.D.-M. were supported by BBSRC grant BB/M018040/1, C.S. and S.D.H. by Wellcome Trust grant 108420/Z/15/Z and J.L.S. by the British Pharmacological Society. We appreciated referee comments that enabled the enhancement of the final version and also the permission of

DOI: 10.1021/acsomega.8b00659
ACS Omega 2018, 3, 7993−8002

ACS Omega to post a preprint of the first submission (https://chemrxiv.org/articles/An_open-access_tool_for_designing_drug_control_into_engineered_proteins/6106541) from which we will ensure a cross-pointer to the final published version.

REFERENCES

(1) Guet, C. C.; Elowitz, M. B.; Hsing, W.; Leibler, S. Combinatorial synthesis of genetic networks. *Science* **2002**, *296*, 1466−1470.

(2) Ro, D. K.; Paradise, E. M.; Ouellet, M.; Fisher, K. J.; Newman, K. L.; Ndungu, J. M.; Ho, K. A.; Eachus, R. A.; Ham, T. S.; Kirby, J.; Chang, M. C.; Withers, S. T.; Shiba, Y.; Sarpong, R.; Keasling, J. D. Production of the antimalarial drug precursor artemisinic acid in engineered yeast. *Nature* **2006**, *440*, 940−943.

(3) Davies, J. A. Synthetic morphology: prospects for engineered, self-constructing anatomies. *J. Anat.* **2008**, *212*, 707−719.

(4) Greber, D.; Fussenegger, M. An engineered mammalian band-pass network. *Nucleic Acids Res.* **2010**, *38*, No. e174.

(5) Cachat, E.; Liu, W.; Hohenstein, P.; Davies, J. A. A library of mammalian effector modules for synthetic morphology. *J. Biol. Eng.* **2014**, *8*, 26.

(6) Morsut, L.; Roybal, K. T.; Xiong, X.; Gordley, R. M.; Coyle, S. M.; Thomson, M.; Lim, W. A. Engineering Customized Cell Sensing and Response Behaviors Using Synthetic Notch Receptors. *Cell* **2016**, *164*, 780−791.

(7) Chan, C. T.; Lee, J. W.; Cameron, D. E.; Bashor, C. J.; Collins, J. J. 'Deadman' and 'Passcode' microbial kill switches for bacterial containment. *Nat. Chem. Biol.* **2016**, *12*, 82−86.

(8) Zhou, X.; Brenner, M. K. Improving the safety of T-Cell therapies using an inducible caspase-9 gene. *Exp. Hematol.* **2016**, *44*, 1013−1019.

(9) Ramos, J. L.; Martínez-Bueno, M.; Molina-Henares, A. J.; Terán, W.; Watanabe, K.; Zhang, X.; Gallegos, M. T.; Brennan, R.; Tobes, R. The TetR family of transcriptional repressors. *Microbiol. Mol. Biol. Rev.* **2005**, *69*, 326−356.

(10) Chen, C. Y.; Ezzeddine, N.; Shyu, A. B. Messenger RNA half-life measurements in mammalian cells. *Methods Enzymol.* **2008**, *448*, 335−357.

(11) Davis, K. M.; Pattanayak, V.; Thompson, D. B.; Zuris, J. A.; Liu, D. R. Small molecule-triggered Cas9 protein with improved genome-editing specificity. *Nat. Chem. Biol.* **2015**, *11*, 316−318.

(12) Nguyen, D. P.; Miyaoka, Y.; Gilbert, L. A.; Mayerl, S. J.; Lee, B. H.; Weissman, J. S.; Conklin, B. R.; Wells, J. A. Ligand-binding domains of nuclear receptors facilitate tight control of split CRISPR activity. *Nat. Commun.* **2016**, *7*, No. 12009.

(13) Chen, X.; Liu, M.; Gilson, M. K. BindingDB: a web-accessible molecular recognition database. *Comb. Chem. High Throughput Screening* **2001**, *4*, 719−725.

(14) Ekins, S.; Clark, A. M.; Southan, C.; Bunin, B. A.; Williams, A. J. Small-molecule Bioactivity Databases. In *High Throughput Screening Methods: Evolution and Refinement*; Bittker, J. A., Ross, N. T., Eds.; Royal Society of Chemistry Publications, 2017.

(15) Gilson, M. K.; Liu, T.; Baitaluk, M.; Nicola, G.; Hwang, L.; Chong, J. BindingDB in 2015: A public database for medicinal chemistry, computational chemistry and systems pharmacology. *Nucleic Acids Res.* **2016**, *44*, D1045−D1053.

(16) Gaulton, A.; Hersey, A.; Nowotka, M.; Bento, A. P.; Chambers, J.; Mendez, D.; Mutowo, P.; Atkinson, F.; Bellis, L. J.; Cibrián-Uhalte, E.; Davies, M.; Dedman, N.; Karlsson, A.; Magariños, M. P.; Overington, J. P.; Papadatos, G.; Smit, I.; Leach, A. R. The ChEMBL database in 2017. *Nucleic Acids Res.* **2017**, *45*, D945−D954.

(17) Berman, H. M.; Burley, S. K.; Kleywegt, G. J.; Markley, J. L.; Nakamura, H.; Velankar, S. The archiving and dissemination of biological structure data. *Curr. Opin. Struct. Biol.* **2016**, *40*, 17−22.

(18) Desaphy, J.; Bret, G.; Rognan, D.; Kellenberger, E. sc-PDB: a 3D-database of ligandable binding sites -10 years on. *Nucleic Acids Res.* **2015**, *43*, D399−D404.

(19) Liu, Z.; Li, Y.; Han, L.; Li, J.; Liu, J.; Zhao, Z.; Nie, W.; Liu, Y.; Wang, R. PDB-wide collection of binding data: current status of the PDBbind database. *Bioinformatics* **2015**, *31*, 405−412.

(20) Harding, S. D.; Sharman, J. L.; Faccenda, E.; Southan, C.; Pawson, A. J.; Ireland, S.; Gray, A. J. G.; Bruce, L.; Alexander, S. P. H.; Anderton, S.; Bryant, C.; Davenport, A. P.; Doerig, C.; Fabbro, D.; Levi-Schaffer, F.; Spedding, M.; Davies, J. A. NC-IUPHAR. The IUPHAR/BPS Guide to PHARMACOLOGY in 2018: updates and expansion to encompass the new guide to IMMUNOPHARMACOLOGY. *Nucleic Acids Res.* **2018**, *46*, D1091−D1106.

(21) Pándy-Szekeres, G.; Munk, C.; Tsonkov, T. M.; Mordalski, S.; Harpsøe, K.; Hauser, A. S.; Bojarski, A. J.; Gloriam, D. E. GPCRdb in 2018: adding GPCR structure models and ligands. *Nucleic Acids Res.* **2018**, *46*, D440−D446.

(22) https://github.com/biasmv/pv.

(23) https://github.com/onursumer/biojs-vis-proteinFeaturesViewer.

(24) Kyte, J.; Doolittle, R. F. A simple method for displaying the hydropathic character of a protein. *J. Mol. Biol.* **1982**, *157*, 105−132.

(25) Southan, C.; Hancock, J. M. A tale of two drug targets: the evolutionary history of BACE1 and BACE2. *Front. Genet.* **2013**, *4*, 293.

(26) Laskowski, R. A.; Jabłońska, J.; Pravda, L.; Vařeková, R. S.; Thornton, J. M. PDBsum: Structural summaries of PDB entries. *Protein Sci.* **2018**, *27*, 129−134.

(27) Kufareva, I.; Ilatovskiy, A. V.; Abagyan, R. Pocketome: an encyclopedia of small-molecule binding sites in 4D. Pocketome: an encyclopedia of small-molecule binding sites in 4D. *Nucleic Acids Res.* **2012**, *40*, D535−D540.

(28) Kelley, L. A.; Mezulis, S.; Yates, C. M.; Wass, M. N.; Sternberg, M. J. The Phyre2 web portal for protein modeling, prediction and analysis. *Nat. Protoc.* **2015**, *10*, 845−58.

(29) Feil, R.; Brocard, J.; Mascrez, B.; LeMeur, M.; Metzger, D.; Chambon, P. Ligand-activated site-specific recombination in mice. *Proc. Natl. Acad. Sci. U.S.A.* **1996**, *93*, 10887−10890.

(30) Kretzschmar, K.; Watt, F. M. Lineage tracing. *Cell* **2012**, *148*, 33−45.

(31) Kumar, M. E.; Bogard, P. E.; Espinoza, F. H.; Menke, D. B.; Kingsley, D. M.; Krasnow, M. A. Mesenchymal cells. Defining a mesenchymal progenitor niche at single-cell resolution. *Science* **2014**, *346*, No. 1258810.

(32) Pippin, J. W.; Kaverina, N. V.; Eng, D. G.; Krofft, R. D.; Glenn, S. T.; Duffield, J. S.; Gross, K. W.; Shankland, S. J. Cells of renin lineage are adult pluripotent progenitors in experimental glomerular disease. *Am. J. Physiol.: Renal. Physiol.* **2015**, *309*, F341−F358.

(33) Lehoczky, J. A.; Robert, B.; Tabin, C. J. Mouse digit tip regeneration is mediated by fate-restricted progenitor cells. *Proc. Natl. Acad. Sci. U.S.A.* **2011**, *108*, 20609−20614.

(34) Zhu, G.; Chow, L. M.; Bayazitov, I. T.; Tong, Y.; Gilbertson, R. J.; Zakharenko, S. S.; Solecki, D. J.; Baker, S. J. Pten deletion causes mTorc1-dependent ectopic neuroblast differentiation without causing uniform migration defects. *Development* **2012**, *139*, 3422−3431.

(35) McFarlane, M. R.; Cantoria, M. J.; Linden, A. G.; January, B. A.; Liang, G.; Engelking, L. J. Scap is required for sterol synthesis and crypt growth in intestinal mucosa. *J. Lipid Res.* **2015**, *56*, 1560−1571.

(36) Haldar, M.; Hedberg, M. L.; Hockin, M. F.; Capecchi, M. R. A CreER-based random induction strategy for modeling translocation-associated sarcomas in mice. *Cancer Res.* **2009**, *69*, 3657−3664.

(37) Rotolo, T.; Smallwood, P. M.; Williams, J.; Nathans, J. Genetically-directed, cell type-specific sparse labeling for the analysis of neuronal morphology. *PLoS One* **2008**, *3*, No. e4099.

(38) Link, K. H.; Shi, Y.; Koh, J. T. Light activated recombination. *J. Am. Chem. Soc.* **2005**, *127*, 13088−13089.

(39) Boutet, A.; De Frutos, C. A.; Maxwell, P. H.; Mayol, J.; Romero, J.; Nieto, M. A. Snail activation disrupts tissue homeostasis and induces fibrosis in the adult kidney. *EMBO J.* **2006**, *25*, 5603−5613.

(40) https://pygtop.readthedocs.io/en/latest.

DOI: 10.1021/acsomega.8b00659
ACS Omega 2018, 3, 7993−8002

(41) https://github.com/samirelanduk/pygtop, https://github.com/samirelanduk/atomium/tree/molecupy1.0.5.
(42) https://github.com/samirelanduk/synpharm.
(43) https://www.postgresql.org/.
(44) https://synpharm.guidetopharmacology.org.

NOTE ADDED IN PROOF

Since GtoPdb has a new release approximately every 2 months the statistics will change from those given in the tables above.

DOI: 10.1021/acsomega.8b00659
ACS Omega 2018, 3, 7993−8002